wCorr Formulas

Paul Bailey

2016-03-04

This document shows the methodology used in computing the correlations in wCorr for the weighted Pearson, Spearman, polyserial, and polychoric correlations.

For the polyserial and polychoric, evidence is offered for the correctness of the methods–including a sketch of a proof of the consistency of both methods and numerical simulations to show:

- The bias of the methods as a function of the true correlation coefficient (ρ) and the number of observations (n)
- The root mean square error (RMSE) of the methods as a function of ρ and n

A second vignette "wCorr Arguments" describes the effect of chaning the `ML` and `fast` arguments.

Methodology

Consider the estimation of the correlation coefficients between two vectors of random variables that are jointly bivariate normal–call the vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$.[1] The i^{th} members of the vectors are then called x_i and y_i and the joint distribution is given by

$$\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix} \sim N\left[\begin{pmatrix}\mu_x\\ \mu_y\end{pmatrix}, \mathbf{\Sigma}\right]$$

where $N(\mathbf{A}, \mathbf{\Sigma})$ is the bivariate normal distribution with mean $\boldsymbol{A}$ and covariance $\mathbf{\Sigma}$.

Pearson and Spearman methodology

The Pearson correlation is computed using the formula

$$\rho_{Pearson} = \frac{\sum_i \left[w_i(x_i - \bar{x}) \times (y_i - \bar{y})\right]}{\sqrt{\sum_i \left(w_i \times (x - \bar{x})^2\right) + \sum_i \left(w_i \times (y - \bar{y})^2\right)}}$$

where $\boldsymbol{W}$ is the vector of weights, $\bar{x} = \frac{1}{\sum w_i} \sum_i w_i x_i$, $\bar{y} = \frac{1}{\sum w_i} \sum_i w_i y_i$, and n is the number of elements in $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{W}$.[2]

The Spearman correlation coefficient is calculated by first taking the rank of the data before the same formula is applied. When data are ranked ties must be handled in some way. The chosen method is to assign the average of all tied ranks. For example, if the second and third rank units are tied then both units would receive a rank of 2.5 (the average of 2 and 3).

[1] The Spearman correlation coefficient can be motivated by a much more broad class of random variables but that is not dealt with in this document.

[2] See the "correlate" function in Stata Corp, Stata Statistical Software: Release 8. College Station, TX: Stata Corp LP, 2003.

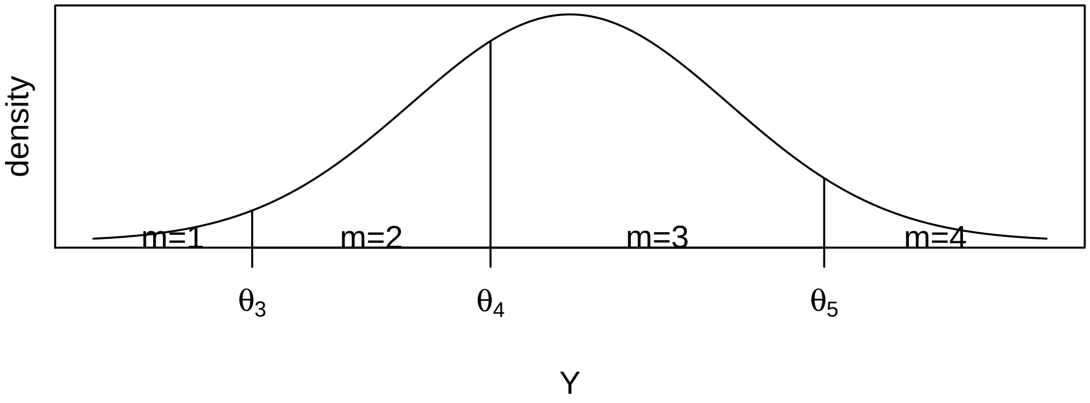
Polyserial methodology

For the polyserial, $\boldsymbol{Y}$ is discretized into a random variable $\boldsymbol{M}$ with t bins according to[3]

$$m_i = \begin{cases} 1 & \text{if} \theta_2 < y_i < \theta_3 \\ 2 & \text{if} \theta_3 < y_i < \theta_4 \\ \vdots & \\ t & \text{if} \theta_{t+1} < y_i < \theta_{t+2} \end{cases}$$

where, for notational convenience, $\theta_2 \equiv -\infty$ and $\theta_{t+2} \equiv \infty$.[4]

The following figure shows the density of $\boldsymbol{Y}$ when $\theta = (-\infty, -2, -0.5, 1.6, \infty)$. Here, for example, any value of $-2 < y_i < -0.5$ ($\theta_3 < y_i < \theta_4$) would have $m_i = 2$.



Notice that μ_y is not identifed (or is irrelevant) because setting $\tilde{\mu}_y = \mu_y + a$ and $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta} + a$ lead to exactly the same values of $\mathbf{M}$ and so one of the two must be artibrary assigned. A convenient decision is to decide $\mu_y \equiv 0$. A similar argument holds for σ_y so that $\sigma_y \equiv 1$.

For $\boldsymbol{X}$ Cox (1974) observes that the MLE mean and standard deviation of $\boldsymbol{X}$ are simply the average and (population) standard deviation of the data and do not depend on the other parameters.[5] This can be taken advantage of by defining $z \equiv \frac{x-\bar{x}}{\hat{\sigma}_x}$.

Combining these simplifications the probability of any given x_i, m_i pair is

$$\Pr\left(\rho = r; Z = z_i, M = m_i\right) = \phi(z_i) \int_{\theta_{m_i+1}}^{\theta_{m_i+2}} dy f(y|Z = z, \rho = r)$$

where $\Pr\left(\rho = r; Z = z, M = m\right)$ is the probability of the event $\rho = r$, $\mathbf{Z} = z$, and $\mathbf{M} = m$, $\phi(\cdot)$ is the standard normal and $f(y|z, \rho)$ is the distribution of y conditional on z and ρ. Because y and z are jointly bivariate normally distributed (by assumption)

$$f(y|Z = z, \rho = r) = N\left(\mu_y + \frac{\sigma_y}{\sigma_z} r(z - \mu_z), (1 - r^2)\sigma_y\right)$$

[3] For a more complete treatment of the polyserial correlation, see Cox, N. R., "Estimation of the Correlation between a Continuous and a Discrete Variable" *Biometrics*, **50** (March), 171-187, 1974.

[4] The indexing is somewhat odd to be consistent with Cox (1974). Nevertheless, this treatment does not use the Cox definition of θ_0, θ_1 or θ_2 which are either not estimated (as is the case for θ_0, and θ_1) or are reappopriated (as is the case for θ_2). Cox calls the correlation coefficient θ_2 while this document uses ρ and uses θ_2 to store $-\infty$ as a convenience so that the vector $\boldsymbol{\theta}$ includes the (infinite) bounds as well as the interior points.

[5] The population standard devaition is used because it is the MLE for the standard deviation.

because both $\boldsymbol{Z}$ and $\boldsymbol{Y}$ are standard normals

$$f(y|Z=z,\rho=r) = N\left(r\cdot z,(1-r^2)\right)$$

Plugging this in

$$\Pr\left(\rho=r,\boldsymbol{\theta};Z=z_i,M=m_i\right) = \phi(z_i)\left[\Phi\left(\frac{\theta_{m_i+2}-r\cdot z_i}{\sqrt{1-r^2}}\right)-\Phi\left(\frac{\theta_{m_i+1}-r\cdot z_i}{\sqrt{1-r^2}}\right)\right]$$

Where $\Phi(\cdot)$ is the standard normal cumulative density function. Using the above probability function as an objective, the log-likelihood is then maximized.

$$\ell(\rho,\boldsymbol{\theta};z,m) = \sum_i w_i \ln\left[\Pr\left(\rho=r,\boldsymbol{\theta};Z=z_i,M=m_i\right)\right]$$

where w_i is the weight of the i^{th} case.

When `ML` is set to `FALSE` the value of the nusiance paramter $\boldsymbol{\theta}$ is chosen to be $\Phi^{-1}(n/N)$ where n is the number of values to the left of the cut point (θ_i value) and N is the number of data points overall. When `ML=TRUE` the joing probability of ρ and $\boldsymbol{\theta}$ is maximized.

Polyserial computation

The derivatives of ℓ can be written out but are not readily computed and so optimization relies only on evaluation of ℓ itself and no derivatives. When the `ML` argumet is set to `FALSE` (the default) a one dimensional optimization of ρ is calculated using `stats::optimize`. When the `ML` argument is set to `TRUE` a multi-dimensional optimization is done for ρ and $\boldsymbol{\theta}$ using `minqa::bobyqa`.

Because the optimization is not perfect when the correlation is in a boundary condition ($\rho\in\{-1,1\}$), a check for perfect correlation is performed before the above optimization by simply seeing if the values of $\boldsymbol{X}$ and $\boldsymbol{M}$ have exactly the same order.

Polychoric methodology

Similar to the polyserial, the polychoric is a simple case of two continuous variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ that have a bivariate normal distribution. In the case of the polyserial the continious (latent) varialbe $\boldsymbol{Y}$ was observed as a discretized variable $\boldsymbol{M}$. For the polychoric this is again true but now the continious (latent) variable $\boldsymbol{X}$ is observed as a discrete variable $\boldsymbol{P}$ in t' bins according to

$$p_i = \begin{cases} 1 & \text{if} \theta'_2 < x_i < \theta'_3 \\ 2 & \text{if} \theta'_3 < x_i < \theta'_4 \\ \vdots \\ t & \text{if} \theta'_{t'+1} < x_i < \theta'_{t'+2} \end{cases}$$

where $\boldsymbol{\theta}$ remains the cut points for the distibution defining the transformation of $\boldsymbol{Y}$ to $\boldsymbol{M}$ and $\boldsymbol{\theta}'$ is the cut points for the tramsofmation from $\boldsymbol{X}$ to $\boldsymbol{P}$. Similar to $\boldsymbol{\theta}$, $\boldsymbol{\theta}'$ has $\theta'_2\equiv-\infty$ and $\theta'_{t'+2}\equiv\infty$.

Similar to in the polyserial, μ_y is not identifed (or is irrelevant) because setting $\tilde{\mu}_y=\mu_y+a$ and $\tilde{\boldsymbol{\theta}}=\boldsymbol{\theta}+a$ lead to exactly the same values of $\mathbf{M}$ and so one of the two must be artibrary assigned. The same is true for μ_x. A convenient decision is to decide $\mu_y=\mu_x\equiv 0$. A similar argument holds for σ_y and σ_x so that $\sigma_y=\sigma_x\equiv 1$.

Then the probability of any given m_i, p_i pair is

$$\Pr\left(\rho = r, \boldsymbol{\theta}, \boldsymbol{\theta}'; P = p_i, M = m_i\right) = \int_{\theta'_{p_i+1}}^{\theta_{p_i+2}} dx \int_{\theta_{m_i+1}}^{\theta_{m_i+2}} dy f(x, y|\rho = r)$$

where ρ is the correlation coefficient.

Using this function as an objective, the log-likelihood is then maximized.

$$\ell(\rho, \boldsymbol{\theta}, \boldsymbol{\theta}'; p, m) = \sum_i w_i \ln\left[\Pr\left(\rho = r, \boldsymbol{\theta}, \boldsymbol{\theta}'; P = p_i, M = m_i\right)\right]$$

Polychoric computation

This again mirrors the treatment of the polyserial. The derivatives of ℓ can be written down but are not readily computed and so when the `ML` argumet is set to `FALSE` (the default) a one dimensional optimization of ρ is calculated using `stats::optimize`. When the `ML` argument is set to `TRUE` a multi-dimensional optimization is done for ρ, $\boldsymbol{\theta}$, and $\boldsymbol{\theta}'$ using `minqa::bobyqa`.

Because the optimization is not perfect when the correlation is in a boundary condition ($\rho \in \{-1, 1\}$), a check for perfect correlation is performed before the above optimization by simply seeing if the values of $\boldsymbol{P}$ and $\boldsymbol{M}$ have a Goodman-Kruskal correlation coefficient of -1 or 1. When this is the case, the MLE of -1 or 1, respectivly, is returned.

Correctness

It is easy to prove the consistency of the $\boldsymbol{\theta}$ for the polyserial and $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ using the non-ML case. Similarly, for ρ, because it is an MLE that can be obtained by taking a derivative and setting it equal to zero, the results are asymtotically unbiased and obtain the Cramer-Rao lower bound.

This does not speak to the small sample properties of these correlation coefficients. Previous work has described their properties by simulation and so that tradition is continued below.

General setup for the unweighted case

A simulation is run several times using the following method:

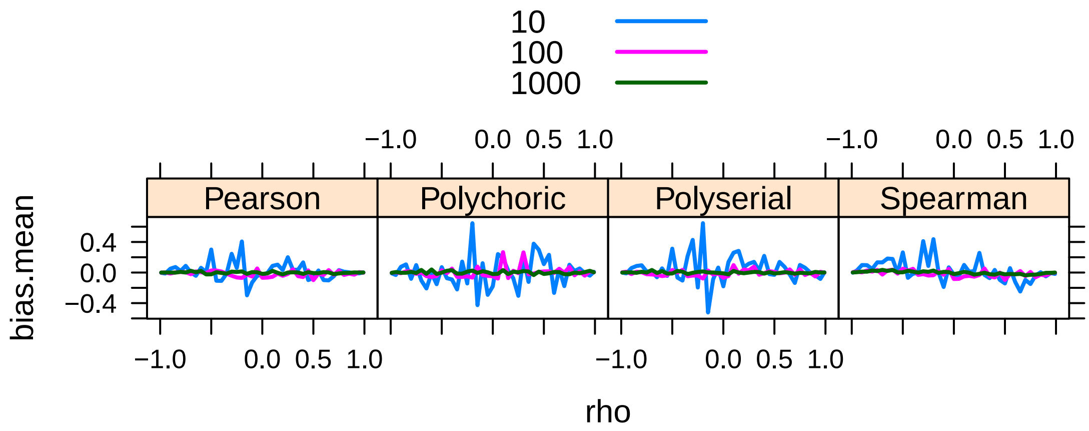
- selecting a value of n (the number of observations)
- selecting a true correlation coefficient ρ
- generating $\boldsymbol{X}$ and $\boldsymbol{Y}$
- selecting the value of t and t' (the number of bins for $\boldsymbol{M}$ and $\boldsymbol{P}$)
- selecting $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$
- confirming that at least 2 levels of $\boldsymbol{M}$ and $\boldsymbol{P}$ are occupied (if not, retrun to generating $\boldsymbol{X}$ and $\boldsymbol{Y}$)
- calculating and recording relevant statistics

bias of the correlations

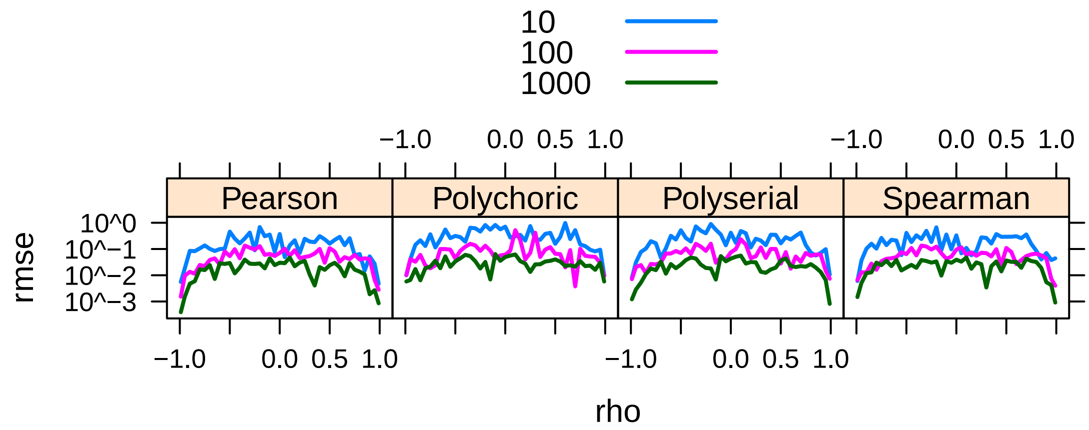
This sections shows the bias of the correlations as a function of the true correlation coefficient, ρ. These simulations were carried out using (junk about true ρ and n grids and number of reps).

RMSE of the correlations as a function of n and ρ

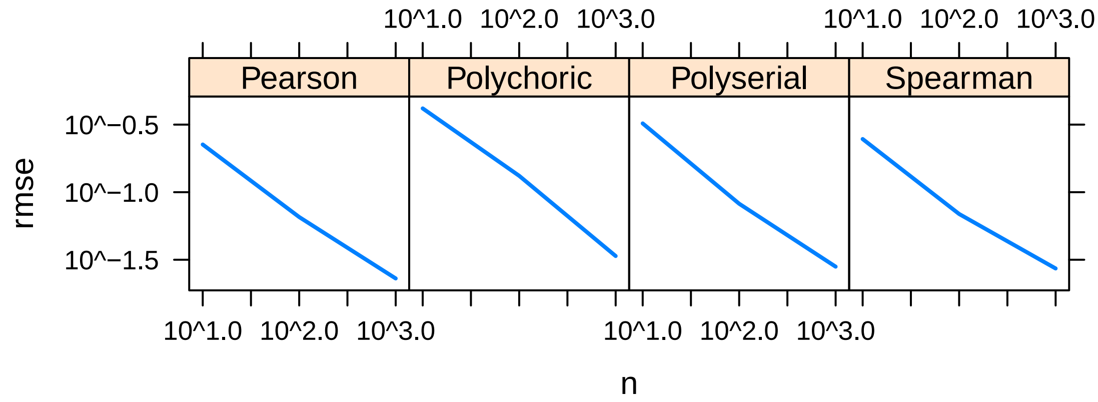
This plot shows the bias as a function of the true correlation ρ



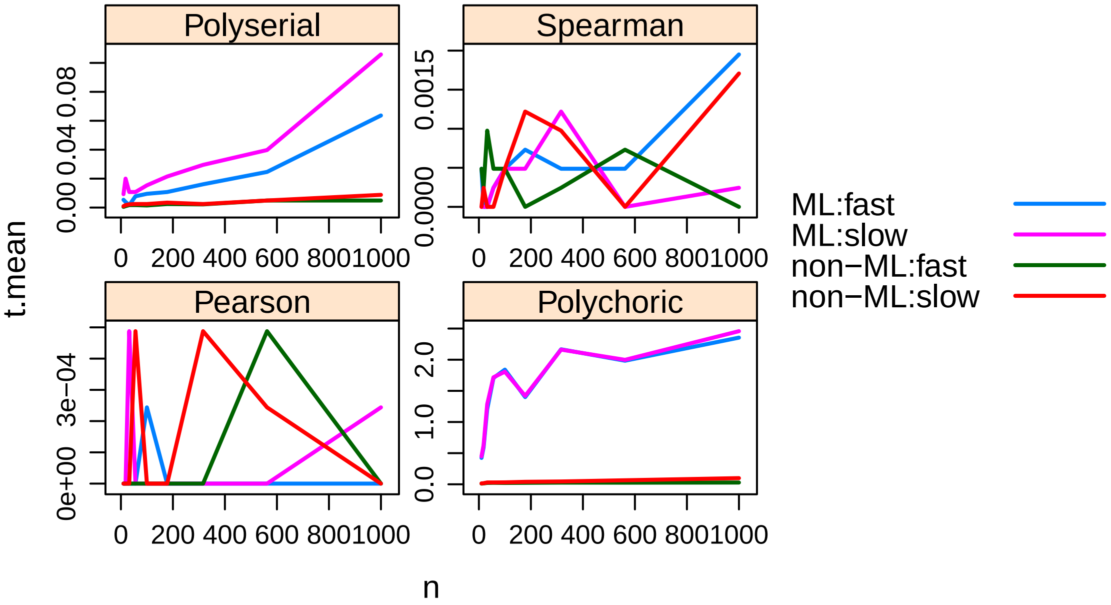
this plot shows the RMSE as a function of ρ



This plot shows the RMSE as a function of n



This plot shows the mean time to compute a single correlation coefficient as a function of ρ by n size.



General setup for the weighted case

For the weighted case the unweighted result would be consistent if there were not something about the higher weight cases that makes them diffeent from the lower weight cases. Thus, while it is not reasonable to always assume that thee is a difference between the high and low weight cases, boht are possible and it serves as a more robust test of the methods in this package to consider a cases where they are associated.

There are many possible ways to associate the higher and lower weight cases. For the purposes of this document it is assumed that there is a third variable $\boldsymbol{W}$ that is correlated with $\boldsymbol{X}$ and $\boldsymbol{Y}$ and that the varible $\boldsymbol{W}$ is also the weight. While in a pratical cases it is unlikely that the weight itself would be causing changes in the relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$, this is a simplifying assumption that is used as a conveniance–averting the need for a fourth variable.

Thus, the general setup is this:

$$\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\\ \mathbf{W}\end{pmatrix} \sim N\left[\begin{pmatrix}\mu_x\\ \mu_y\\ \mu_w\end{pmatrix}, \Sigma\right]$$

for conveniance, and without loss of generality, it is assumed that $\mu_x = \mu_y = \mu_w \equiv 0$ and that

$$\Sigma \equiv \begin{pmatrix}1 & \rho & a\\ \rho & 1 & b\\ a & b & 1\end{pmatrix}$$

where ρ remails the parameter of interest and $a \in [0,1]$ and $b \in [0,1]$ are the correlation between $\boldsymbol{X}$ and $\boldsymbol{W}$ and the correlation between $\boldsymbol{Y}$ and $\boldsymbol{W}$, respectively.

Simulations are carred out in the same fashion as before, except that after ρ is elected, a and b are aslo selected.

Polyserial simulation study, weighted

- method(s) of selecting n, ρ, ρ, a, b, t, t', $\boldsymbol{\theta}$, $\boldsymbol{\theta}'$.

then make these plots (note: include Pearson and Spearman): * Compare weighted and unweighted bias as a function of ρ and $a = b$, all $n =$ something reasonable based on the final bias chart . * Compare weighted and unweighted RMSE as a function of ρ and $a = b$, all $n =$ something reasonable based on the final bias chart.

- Compare weighted and unweighted bias as a function of ρ and a while $b = 0$, all $n =$ something reasonable based on the final bias chart.
- Compare weighted and unweighted RMSE as a function of ρ and a while $b = 0$, all $n =$ something reasonable based on the final bias chart.

Polychoric simulation study, weighted